Customizing Clustering algorithm for Data mining in CRM

E.Manigandan, Research Scholar, SCSVMV
University, Enathur, Kanchipuram-631561,
ibmmani78@yahoo.com

Dr.V.Shanthi, Professor, Dept. of MCA., St.
Joseph's College of Engineering, Chennai -119.
drvshanthi@yahoo.co.in

Magesh Kasthuri, Research Scholar, SCSVMV
University, Enathur, Kanchipuram-631561,
magesh.kasthuri@wipro.com

Abstract— **Spectral Clustering closely works with Customer by collecting various information from customer and prepare a relation among these data to provide a meaningful relation of information. This can be used further to predict market trend and key expectation in marketing improvisation. Hence this algorithm is one of the key research topics in the area of Customer Relationship Management (CRM) for mining customer data and prepares informative reports.**

Keywords-– K-means, Spectral clustering, Predictive model, Data mining, Data analysis, Big data

I. INTRODUCTION

Data mining is the process of discovering meaningful new correlations, patterns and trends by sifting through large amounts of data stored in repositories, using pattern recognition technologies as well as statistical and mathematical techniques. It involves scientific and statistical approach of the exploration and analysis, by automatic or semiautomatic means, of large quantities of data in order to discover meaningful patterns and rules.

While new to many industries, predictive analytics is a proven technology that's been used successfully for many decades. It encompasses a variety of techniques that derive insights from data with one clear-cut goal: find the best action to a given situation. Apart from predictive analysis which is based on historical results, there are other techniques in Data mining and several algorithms implemented in Data mining concepts for better data analysis. Spectral clustering is one of such mechanism which is more popular in social networking based marketing field to predict market demands and improvise the area of marketing research and in turn profit.

In this paper we are going to discuss, some of the key algorithms including K-means and Spectral clustering and also discuss some of the improvisation on Spectral clustering which can benefit data mining to provide more accurate and predictive results in the field of CRM.

II. GROWTH OF DATA MINING RESEARCH

Data mining is a classical methodology for analysing raw data collected in various methods and forms. There are two types of data mining models – Predictive and Descriptive. Predictive analysis is based on previously collected and analysed information whereas descriptive is a free flow analysis of information without any pre-setup of data results.

In a typical example of data warehousing platform, where there are millions of data collected and kept in unsorted fashion, data mining is a trivial job and there are many preparatory steps involved in such large dataset analysis. For example, Data preparation for Data mining may be part of the Data Warehousing where Data Warehouse not a requirement for Data Mining. Classical statistics based on elegant theory and restrictive data assumptions are fine if data sets are small and assumptions met for the scope and boundary of data analysis. There are various tools for such data analysis and in newer tools; pattern finding is data-driven rather than user-driven.

The key ingredient of a data mining process can be classified into computing power of the system we use, statistical and learning algorithm we apply for data analysis and improved data collection and management of data processed. This is depicted in below diagram

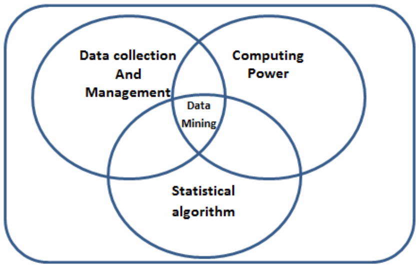


Figure 1: Data Mining principles

III. DATA ANALYTICS

Data analytics is the field of data mining where we apply scientific approaches in analysis data and predicting or computing results of interest in terms of reports or processed data or trend analysis etc., In a modern big data analysis world, there many types of data analytics and it is very important to understand the key processing technique of data analytics in order to design an efficient algorithm for data mining and analysis.

Classical Data Analytics – It is based on Conversion optimization and analysis eg: site or advertisement or goal optimization using segmentation, targeting, Multivariate Testing/Analysis to name a few. It is useful in Marketing Campaign Analysis & Optimization and Integration with trend analysis report.

Advanced Data Analytics - CRM integration for enhanced targeting and cross-sell or up-sell insights and dis-engagement setups and it is used in Propensity Modelling of anonymous data analysis and customers information processing. It also helps in Behavioural segmentation of anonymous users using audience targeting data from publishers. It is used in Multi-channel marketing optimization and Cohort Analysis.

Experience Optimization – This technique is used in Measurement plan & implementation for optimizing customer pain areas and delight opportunities to improvise marketing strategy. Business impact quantification and real-time action triggers for usability and other experience obstacles can be easily handled in this technique. It helps in Personal development and customer journey mapping.

Social Analytics - Social Analytics is the modern day trend analysis in new age social media field (eg: Twitter data analysis, Facebook usage statistics etc.,) and it is based on Social Intelligence Program Setup like influencer analysis and social CRM. Return of Investment (ROI) is the key target of this analytics tool and it helps in 360-degree multichannel customer view setup with Special need competitive intelligence and digital behavioural segmentation in multivariate testing and analysis

IV. TECHNIQUES IN DATA MINING

There are several techniques involved in data analytics in a data mining world where the choice of technique depends on various parameters including need for analysis, ROI, data density to name a few. For example, there are a broad range of predictive analytics techniques are available in modern data analysis and some key techniques extensively used in the industry are listed below. The scope of this paper is not to analyse these techniques but to get the key technique and improvise it (eg: Clustering).

- Logistic regression
- Linear regression
- Discriminant analysis
- Scorecards
- Decision trees
- Clustering
- Neural networks
- Survival analysis
- Market basket analysis
- Collaborative filtering
- Support vector machines
- Discrete choice modeling
- Causal models

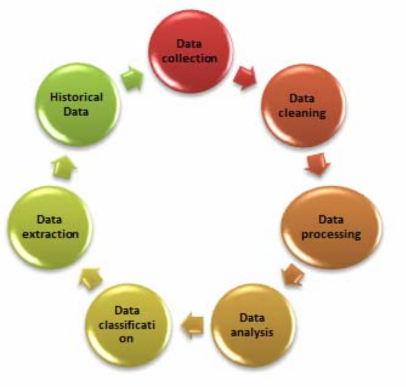


Figure 2: Life cycle of Data mining

V. DATA PREPARATION

When preparing the environment for such data analysis, it is important step and preliminary step in the process in identifying the right scope of data to be sourced for data analysis. This means, that we need to choose right mechanism for data collection and Data sources can be company specific internal data or external data from various sources. For example, Some key data used in Insurance industry are:

- Demographic data
- Credit history from bureaus
- Transaction history
- Payment history
- Life stage data
- Collection contacts
- Customer service contacts
- Channel usage
- Web/Social media data
- Customer survey data

Then we need to encounter the step of choosing right tool for data analysis. There are several tools available for data mining, analysis, visualization, predictive modeling and optimization. Some key tools used in the industry are SAS, SPSS, Stata, R,IBM Intelligent Miner, FICO Model Builder, Rapid miner, ILOG, WEKA to name a few.

VI. EXPERIMENTAL SETUP

When analyzing data for clustering, we need to find a relation between data and form groups among them. This is based on various techniques as shown below:

Conditional logic – For example, when we are analysing data on health drink consumption, we need to analyse the consumer nature (athlete, student, working class people) and the age range (<30 age, 30-45 age, 46-60 age people etc.,)

Association – For example, when there is an increase in sale of a consumer product (eg: Toothpaste), there is equal amount of sale growth in another consumer product (eg: tooth brush)

Trends and Variations – Sale of a particular product increase in summer and decreases in winter (eg: Sun cream or cool drinks).

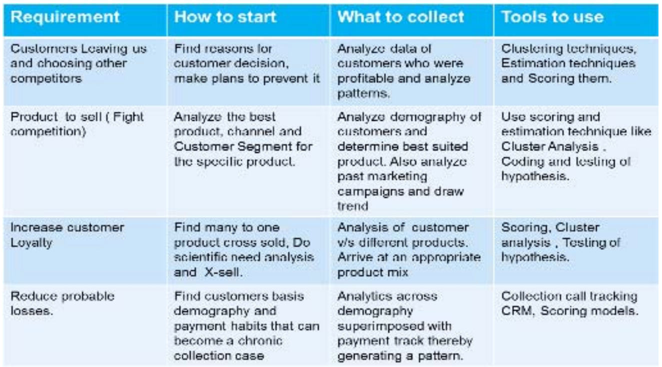
Outcome prediction – When a campaign runs on a particular product in a shopping mall, how many users respond to that and shows interest in the product.

Forecasting – Based on sales growth of 3 quarters in a year, what is the calculative and predictive sale of last quarter of the year.

Deviation detection – How much products are not trust worthy in the market where consumer doesn't have trust in the brand.

Link analysis – When there is a job hunt, there is likely to increase the demand of bank account openings and credit card applications.

TABLE I. CASE STUDY OF INSURANCE DATA PROCESSING



Requirement	How to start	What to collect	Tools to use
Customers Leaving us and choosing other competitors	Find reasons for customer decision, make plans to prevent it	Analyze data of customers who were profitable and analyze patterns.	Clustering techniques, Estimation techniques and Scoring them.
Product to sell (Fight competition)	Analyze the best product, channel and Customer Segment for the specific product.	Analyze demography of customers and determine best suited product. Also analyze past marketing campaigns and draw trend	Use scoring and estimation technique like Cluster Analysis . Coding and testing of hypothesis.
Increase customer Loyalty	Find many to one product cross sold, Do scientific need analysis and X-sell.	Analysis of customer v/s different products. Arrive at an appropriate product mix	Scoring, Cluster analysis , Testing of hypothesis.
Reduce probable losses.	Find customers basis demography and payment habits that can become a chronic collection case	Analytics across demography superimposed with payment track thereby generating a pattern.	Collection call tracking CRM, Scoring models.

VII. IMPROVISATION IN CLUSTERING ALGORITHM

Some of the key advantages of clustering algorithm over K-means algorithm are that it is based on Models which can be built very quickly and Suitable for large data sets. They are easy to understand and give reasons for a decision taken when doing data analysis. Clustering analysis shows trend in handling non-numeric data very well than K-means as it requires Minimum amount of data transformation.

Though there are some Limitations also there in Clustering. For example, it leads to an artificial sense of clarity and the trees of relation built in Clustering left to grow without bound and sometimes take longer to build and become unintelligible and hence it may over fit the data which we feed for analysis.

In a simple example of CRM analysis on a retail industry, the volume of data is exploding, with data stores in different locations and in varying and often proprietary systems. Combined with significant data processing activity and changing regulations, the result is stale, un-normalized and unconsolidated data that poses a serious risk of financial loss. The need to deploy a high performance platform for the purpose of market data consolidation and analytics is now more important than ever. Analytics cannot begin until the data is loaded into the system. This is the number one overlooked metric when considering such a data warehousing system, but its implications are felt throughout the process.

Spectral clustering is most popular data mining technique for big data analysis particularly in the field of social computing and later introduced in various other fields like medical science, customer relationship management (CRM), retail stores, manufacturing and health care. Clustering nodes in a graph is a useful general technique in data mining of large network data sets.

This is where there is improvisation required in Clustering where we introduce clustering based on

Five Basic Factors of the data set viz attribute type, attribute classification, attribute value range, attribute uniqueness and attribute filtering possibility. This Customized algorithm is named as M-Clustering algorithm where M defines the Five Means of training classification.

SUMMARY

This custom clustering enables to maintain a store where the data model can be stored in our own clustering representation in a matrix format associating each unique field in an indexed searchable fashion. This helps in quickly process data without re-processing it when there is a growth in data on top of existing data. This delta (change in data growth) of record changes will be taken for further analysis and it runs in unsupervised methodology by storing various fields in indexed model and based on the result required, the field selection and its corresponding test set selection would happen. M-Cluster algorithm can be customized in such a way that it can run the training set cycle multiple times (and hence cost efficient as the cycle runs in a random selected smaller data units only) to refine the result ratio and error rate.

Since M-Cluster algorithm is using ranking based samples as compared to Best first based search in K-Means or other most commonly used data mining algorithms, there is a definite benefit in M-Cluster when used with continuous growing set of data and also in improvised time taken for processing the records.

REFERENCES

[1] Ng, Andrew Y., Michael I. Jordan, and Yair Weiss. "On spectral clustering: Analysis and an algorithm." Advances in neural information processing systems 2 (2002): 849-856.

[2] Von Luxburg, Ulrike. "A tutorial on spectral clustering." Statistics and computing 17.4 (2007): 395-416.

[3] Zelnik-Manor, Lihi, and Pietro Perona. "Self-tuning spectral clustering." Advances in neural information processing systems. 2004.

[4] Dhillon, Inderjit S., Yuqiang Guan, and Brian Kulis. "Kernel k-means: spectral clustering and normalized cuts." Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2004.

[5] Jordan, F. R. B. M. I., and F. Bach. "Learning spectral clustering." Adv. Neural Inf. Process. Syst 16 (2004): 305-312.

[6] Verma, Deepak, and Marina Meila. "A comparison of spectral clustering algorithms." University of Washington Tech Rep UWCSE030501 1 (2003): 1-18.

[7] Dhillon, Inderjit S., Yuqiang Guan, and Brian Kulis. A unified view of kernel k-means, spectral clustering and graph cuts. Computer Science Department, University of Texas at Austin, 2004.

[8] Tang, Lei, and Huan Liu. "Leveraging social media networks for classification." Data Mining and Knowledge Discovery 23.3 (2011): 447-478.

[9] Lazer, David, et al. "Life in the network: the coming age of computational social science." Science (New York, NY) 323.5915 (2009): 721.

[10] Gundecha, Pritam, and Huan Liu. "Mining social media: a brief introduction." Tutorials in Operations Research 1.4 (2012).

[11] Kameshwaran, K., and K. Malarvizhi. "Survey on Clustering Techniques in Data Mining." International Journal of Computer Science and Information Technologies 5.2 (2014): 2272-2276.

[12] Kumar, Narander, Vishal Verma, and Vipin Saxena. "Cluster Analysis in Data Mining using K-Means Method." International Journal of Computer Applications 76.12 (2013).

[13] Gelman, Andrew, et al. Bayesian data analysis. Vol. 2. Boca Raton, FL, USA: Chapman & Hall/CRC, 2014.

[14] Jirapech-Umpai, Thanyaluk, and Stuart Aitken. "Feature selection and classification for microarray data analysis: Evolutionary methods for identifying predictive genes." BMC bioinformatics 6.1 (2005): 148.

[15] Gelman, Andrew, and Jennifer Hill. Data analysis using regression and multilevel/hierarchical models. Cambridge University Press, 2006.

AUTHORS PROFILE

E.Manigandan is a Professor in Sri Sankara Arts and Science College, Enathur and doing his research study on Data mining in CRM.

Dr.V.Shanthi is professor in Dept. of Computer Applications, St. Joseph’s college of Engineering, Chennai-119

Magesh Kasthuri is a research scholar and a Senior Technical Consultant in Wipro Technologies, Bangalore.